Z. Wahrscheinlichkeitstheorie verw. Geb. 24, 279–286 (1972)

On Sampling from a Finite Set of Independent Random Variables

Bengt von Bahr

0. Let $X_1, X_2, \ldots, X_N$ be a set of N independent random variables and let S_n be the sum of n ($\leq N$) of them chosen at random. In this paper we will show that S_n is for large values of n and under certain mild conditions on the distributions of the X_k's, approximately Gaussian, and we will give an estimate of the remainder term.

This general situation covers for example the case of two-stage sampling. Bikelis [1] has proved a result similar to the present one for simple random sampling, which, of course, can be regarded as a special case of two stage sampling. Bikelis uses an expression for the characteristic function of the sum S_n given by Erdös and Rényi [2], which, as far as I can see, can not be used in the present more general situation.

In the case $n=N$, the sum S_n simply is the sum of n independent random variables, and then the remainder term given here essentially coincides with the classical ones given by Esséen [3].

1. In order to define the sampling procedure exactly, we introduce a random indicator vector $\mathbf{I}=(I_1, I_2, \ldots, I_N)$, where $I_k=0$ or 1, $1 \leq k \leq N$, such that S_n contains the term X_k if and only if $I_k=1$. We now have $S_n=\sum_{k=1}^{N} I_k X_k$. $\mathbf{I}$ is assumed to be independent of the set $X_1, X_2, \ldots, X_N$, and for every ordered sequence $\mathbf{i}=(i_1, i_2, \ldots, i_N)$ of n ones and $N-n$ zeros, we put $P(\mathbf{I}=\mathbf{i})=1\Big/\binom{N}{n}$. We have $EI_k=\frac{n}{N}=f=$ the sampling ratio, and $EI_k I_l=\frac{n}{N}\cdot\frac{n-1}{N-1}$ for $k \neq l$. We introduce the moments $EX_k=\mu_k$ and $EX_k^2=\beta_k$ and then get

$$ES_n=\sum_{k=1}^{N} EI_k X_k = f\sum_{k=1}^{N}\mu_k$$

$$ES_n^2=\frac{n}{N}\sum_k \beta_k+\frac{n}{N}\,\frac{n-1}{N-1}\sum_{k\neq l}\mu_k\mu_l.$$

We will now assume that the scale is chosen so that

$$\sum_{k=1}^{N}\mu_k=0, \tag{1.1}$$

$$\frac{1}{N}\sum_{k=1}^{N}\beta_k=1. \tag{1.2}$$

We then get $ES_n=0$ and

$$\operatorname{Var} S_n = n\left(1-\frac{n-1}{N-1}\alpha^2\right) \qquad \text{where } \alpha^2=\frac{1}{N}\sum_{k=1}^{N}\mu_k^2.$$

In the following sections we will prove that $S_n/\sqrt{n}$ is approximately Gaussian with zero mean and variance $1-f\alpha^2$, and also give an estimate of the remainder term.

2. In this section we will give an exact expression for the characteristic function $\bar{f}_n(t)$ of $S_n/\sqrt{n}$, which is suitable for estimations of the remainder term.

Let $f_k(t)$ be the characteristic function of X_k, $k=1,k,\ldots,N$. Then

$$\bar{f}_n(t)=\binom{N}{n}^{-1}\sum\left(\prod_j f_j\left(\frac{t}{\sqrt{n}}\right)\right)$$

where the products are taken over n different indices j, and the sum is taken over all $\binom{N}{n}$ different combinations of those indices. We now multiply both sides with the same function $e^{t^2/2}$ and get

$$e^{t^2/2}\bar{f}_n(t)=\binom{N}{n}^{-1}\sum\left(\prod_j e^{t^2/2n} f_j\left(\frac{t}{\sqrt{n}}\right)\right).$$

Now, if $g(z)$ is a function of a complex variable z, analytic in a neighbourhood of the origin, we will denote by $\{g(z)\}_n$ the coefficient of z^n in its MacLaurin series. We may thus write

$$e^{t^2/2}\bar{f}_n(t)=\binom{N}{n}^{-1}\left\{\prod_{k=1}^{N}\left(1+z\,e^{t^2/2n} f_k\left(\frac{t}{\sqrt{n}}\right)\right)\right\}_n.$$

Introducing the functions

$$b_k(t)=e^{t^2/2n} f_k\left(\frac{t}{\sqrt{n}}\right)-1, \qquad k=1,2,\ldots,N \tag{2.1}$$

and

$$B_j(t)=\frac{(-1)^{j+1}}{j}\sum_{k=1}^{N} b_k^j(t), \qquad j=1,2,\ldots,n \tag{2.2}$$

we then get

$$\begin{aligned}
e^{t^2/2}\bar{f}_n(t)&=\binom{N}{n}^{-1}\left\{(1+z)^N\prod_{k=1}^{N}\left(1+\frac{z}{1+z}b_k(t)\right)\right\}_n\\
&=\binom{N}{n}^{-1}\left\{(1+z)^N\exp\left(\sum_{k=1}^{N}\log\left(1+\frac{z}{1+z}b_k(t)\right)\right)\right\}_n\\
&=\binom{N}{n}^{-1}\left\{(1+z)^N\exp\left(\sum_{k=1}^{N}\sum_{j=1}^{\infty}\frac{(-1)^{j+1}}{j}\left(\frac{z}{1+z}\right)^j b_k^j(t)\right)\right\}_n\\
&=\binom{N}{n}^{-1}\left\{(1+z)^N\exp\left(\sum_{j=1}^{n}\left(\frac{z}{1+z}\right)^j B_j(t)\right)\right\}_n
\end{aligned}$$

where the summation over j has been restricted to $j \leqq n$, since greater values of j do not contribute to the coefficient for z^n. By expanding the different exponential functions in separate power series of their arguments and multiplying all series together, we get $(B_j = B_j(t))$:

$$\begin{aligned} e^{t^2/2} \bar{f}_n(t) &= \binom{N}{n}^{-1} \left\{ (1+z)^N \sum \left(\frac{z}{1+z}\right)^{i_1} \frac{B_1^{i_1}}{i_1!} \left(\frac{z}{1+z}\right)^{2 i_2} \frac{B_2^{i_2}}{i_2!} \cdots \left(\frac{z}{1+z}\right)^{n i_n} \frac{B_n^{i_n}}{i_n!} \right\}_n \\ &= \binom{N}{n}^{-1} \sum \frac{B_1^{i_1} B_2^{i_2} \ldots B_n^{i_n}}{i_1!\, i_2! \ldots i_n!} \{ z^{i_1 + 2 i_2 + \cdots + n i_n} (1+z)^{N - i_1 - 2 i_2 - \cdots - n i_n} \}_n \\ &= \binom{N}{n}^{-1} \sum \frac{B_1^{i_1} B_2^{i_2} \ldots B_n^{i_n}}{i_1!\, i_2! \ldots i_n!} \binom{N - i_1 - 2 i_2 - \cdots - n\, i_n}{n - i_1 - 2 i_2 - \cdots - n\, i_n} \end{aligned}$$

where the summations are taken over all combinations of integers $i_j \geqq 0$, $1 \leqq j \leqq n$, satisfying $i_1 + 2 i_2 + \cdots + n\, i_n \leqq n$.

We will now examine the binomial coefficients. Putting $\sum_{j=1}^{n} j\, i_j = p$, we have

$$\binom{N}{n}^{-1} \binom{N-p}{n-p} = \left(\frac{n}{N}\right)^p C_{N,n,p} \tag{2.3}$$

where

$$C_{N,n,p} = \frac{\left(1 - \frac{1}{n}\right) \cdots \left(1 - \frac{p-1}{n}\right)}{\left(1 - \frac{1}{N}\right) \cdots \left(1 - \frac{p-1}{N}\right)}. \tag{2.4}$$

If we define $C_{N,n,p} = 0$ when $p > n$, we obtain the following final expression

$$e^{t^2/2} \bar{f}_n(t) = \sum_{\substack{i_j \geqq 0 \\ 1 \leqq j \leqq n}} \frac{(f B_1)^{i_1} (f^2 B_2)^{i_2} \ldots (f^n B_n)^{i_n}}{i_1!\, i_2! \ldots i_n!} C_{N, n, \sum_{j=1}^{n} j i_j}. \tag{2.5}$$

In the following section we shall expand the functions $B_j(t)$ in power series of t. These series will reveal that for large values of n, $f^j B_j(t)$ is small for $j \neq 2$ but $f^2 B_2(t) \approx f \alpha^2 t^2/2$. These facts, together with the fact that $C_{N,n,p} \approx 1$ when n is large, will give that

$$e^{t^2/2} \bar{f}_n(t) \approx \sum \frac{(f^2 B_2)^{i_2}}{i_2!} \approx e^{f \alpha^2 t^2/2}.$$

3. We now assume that all moments of the third order are finite, and put $E|X_k|^3 = \gamma_k$ and $\max_{1 \leqq k \leqq N} \gamma_k = \gamma$. We will denote by θ unspecified functions satisfying $|\theta| \leqq 1$, and assume that t satisfies

$$0 \leqq t \leqq \frac{1 - f \alpha^2}{10 \gamma} \sqrt{n}. \tag{3.1}$$

From the general inequalities for moments, we get

$$|\mu_k| \leqq \beta_k^{\frac{1}{2}} \leqq \gamma_k^{\frac{1}{3}}$$

and thus

$$\frac{1}{N}\sum_{k=1}^{N}\mu_k^2 \leqq \frac{1}{N}\sum_{k=1}^{N}\beta_k \leqq \frac{1}{N}\sum_{k=1}^{N}\gamma_k^{\frac{2}{3}} \leqq \gamma^{\frac{2}{3}}$$

which gives $\alpha^2 \leqq 1 \leqq \gamma$.

Thus (3.1) implies $0 \leqq t/\sqrt{n} \leqq 0.1$. We now get

$$e^{t^2/2n} = 1 + \frac{t^2}{2n} + 0.51\,\theta\left(\frac{t^2}{2n}\right)^2 = 1 + \frac{t^2}{2n} + \frac{\theta\, t^3}{70\, n\sqrt{n}}$$

and from (2.1)

$$b_k(t) = \left(1 + \frac{t^2}{2n} + \frac{\theta\, t^3}{70\, n\sqrt{n}}\right) f_k\left(\frac{t}{\sqrt{n}}\right) - 1$$

$$= f_k\left(\frac{t}{\sqrt{n}}\right) - 1 + \frac{t^2}{2n} f_k\left(\frac{t}{\sqrt{n}}\right) + \frac{\theta\, t^3}{70\, n\sqrt{n}}.$$

By using the usual Taylor expansion for a characteristic function,

$$f_k(t) = 1 + i\, t\, \mu_k - \frac{t^2}{2}\beta_k + \frac{\theta}{6} t^3 \gamma_k$$

we get

$$b_k(t) = \frac{i\, t\, \mu_k}{\sqrt{n}} + \frac{(1-\beta_k)\, t^2}{2n} + 0.7\,\theta\frac{\gamma\, t^3}{n\sqrt{n}}. \tag{3.2}$$

Introducing the inequality (3.1) in the last terms, we also have

$$b_k(t) = \frac{i\, t\, \mu_k}{\sqrt{n}} + 0.57\,\theta\frac{t^2}{n}\gamma^{\frac{2}{3}} \tag{3.3}$$

and

$$b_k(t) = 1.1\,\theta\frac{\gamma^{\frac{1}{3}}\, t}{\sqrt{n}}. \tag{3.4}$$

From (3.3) we get

$$b_k^2(t) = -\frac{\mu_k^2\, t^2}{n} + 1.2\,\theta\frac{t^3\,\gamma}{n\sqrt{n}}. \tag{3.5}$$

We will now use these expressions and (2.2) to estimate $f^j B_j(t)$. From (3.2) we get, taking into account that $\sum_{k=1}^{N}\mu_k = \sum_{k=1}^{N}(1-\beta_k) = 0$

$$f B_1(t) = 0.7\,\theta\frac{\gamma\, t^3}{\sqrt{n}} \tag{3.6}$$

from (3.5) we get

$$f^2 B_2(t) = \frac{f\alpha^2\, t^2}{2} + 0.6\,\theta f\frac{\gamma\, t^3}{\sqrt{n}} \tag{3.7}$$

and from (3.4)

$$f^j B_j(t) = N\theta\left(1.1 f\frac{\gamma^{\frac{1}{3}}\, t}{\sqrt{n}}\right)^j, \quad j = 1, 2, \ldots. \tag{3.8}$$

We now arrive at three expressions, which we will use in the next section: (3.6) and (3.8) give

$$f B_1 + \sum_{j=3}^{n} f^j B_j(t) = 2.2\,\theta \frac{\gamma t^3}{\sqrt{n}} \tag{3.9}$$

(3.7) and (3.1) give

$$f^2 B_2(t) = f\alpha^2 t^2/2 + 0.06\,\theta\, t^2 (1 - f\alpha^2) = 0.6\,\theta\, t^2 = 0.06\,\theta\, t \sqrt{n} \tag{3.10}$$

and (3.1) and (3.10) give

$$\sum_{j=1}^{n} f^j B_j(t) = f\alpha^2 t^2/2 + 0.28\,\theta\, t^2 (1 - f\alpha^2). \tag{3.11}$$

We also must examine the difference between $C_{N,n,p}$ and 1. From (2.4) it at once follows $0 \leqq C_{N,n,p} \leqq 1$. It is also immediately clear that

$$C_{N,n,p} \geqq \left(1 - \frac{1}{n}\right) \dots \left(1 - \frac{p-1}{n}\right).$$

Now, $1 - x \geqq \exp(-2x)$ for $0 \leqq x \leqq \frac{1}{2}$, so when $p \leqq n/2$, we have

$$C_{N,n,p} \geqq \exp\left(-2\left(\frac{1}{n} + \cdots + \frac{p-1}{n}\right)\right) = \exp(-p(p-1)/n) \geqq 1 - p(p-1)/n \geqq 1 - p^2/n.$$

But this inequality is always true for $p \geqq n/2$ if $n \geqq 4$, which we will always assume, and thus for all p:

$$1 - \frac{p^2}{n} \leqq C_{N,n,p} \leqq 1. \tag{3.12}$$

4. We now return to (2.5) and divide the sum $\sum$ of the right side into two parts $\sum = \sum_1 + \sum_2$, where $\sum_1$ is the sum of all terms with $i_1 = i_3 = i_4 = \cdots = i_n = 0$ and $\sum_2$ is the rest. Let $C_i(t)$ be functions, obtained in the preceeding section, satisfying $|f^j B_j(t)| \leqq C_j(t)$. We can then majorize $|\sum_2|$ by replacing every $f^j B_j(t)$ by $C_j(t)$ and $C_{N,n,p}$ by 1. We then obtain $(C_j = C_j(t))$:

$$\left|\sum\nolimits_2\right| \leqq \exp(C_1 + C_2 + \cdots + C_n) - e^{C_2} = e^{C_2}\big(\exp(C_1 + C_3 + \cdots + C_n) - 1\big)$$

By using the elementary inequality

$$|e^x - 1| \leqq |x|\, e^{|x|} \tag{4.1}$$

we get $|\sum_2| \leqq (C_1 + C_3 + \cdots + C_n) \exp(C_1 + C_2 + C_3 + \cdots + C_n)$. (3.9) and (3.11) now give

$$\left|\sum\nolimits_2\right| \leqq \frac{2.2\gamma t^3}{\sqrt{n}} \exp\big(f\alpha^2 t^2/2 + 0.28\, t^2 (1 - f\alpha^2)\big)$$

that is

$$\left|\exp(-t^2/2) \sum\nolimits_2\right| \leqq \frac{2.2\gamma t^3}{\sqrt{n}} \exp\big(-0.22\, t^2 (1 - f\alpha^2)\big). \tag{4.2}$$

We will now estimate $\sum_1$. By definition

$$\sum\nolimits_1 = \sum_{i=0}^{\infty} \frac{(f^2 B_2)^i}{i!} C_{N,n,2i} = \exp(f^2 B_2) - \sum_{i=0}^{\infty} \frac{(f^2 B_2)^i}{i!} (1 - C_{N,n,2i}). \tag{4.3}$$

From (3.7), (4.1) and (3.10) we immediately get

$$|\exp(f^2 B_2)-\exp(f\alpha^2 t^2/2)| \leqq 0.6 f \frac{\gamma t^3}{n} \exp\left(f\alpha^2 t^2/2+0.06 t^2(1-f\alpha^2)\right)$$

that is

$$\left|\exp(-t^2/2+f^2 B_2)-\exp\left((1-f\alpha^2) t^2/2\right)\right| \leqq 0.6 f \frac{\gamma t^3}{\sqrt{n}} \exp\left(-0.44 t^2(1-f\alpha^2)\right). \quad (4.4)$$

By using (3.12) we obtain for the last sum $\sum_3$ in (4.3):

$$\left|\sum\nolimits_3\right| \leqq \sum_{i=0}^{\infty} \frac{C_2^i}{i!} \frac{(2i)^2}{n} = \frac{4}{n} C_2(1+C_2) e^{C_2}$$

and from (3.10)

$$\left|\exp(-t^2/2)\sum\nolimits_3\right| \leqq 0.24 \frac{t(1+t^2)}{\sqrt{n}} \exp\left(-0.44(1-f\alpha^2) t^2\right). \quad (4.5)$$

Combining the expressions (4.2), (4.4) and (4.5), we now obtain

$$\left|\tilde{f}_n(t)-\exp\left(-(1-f\alpha^2) t^2/2\right)\right| \leqq 3.1 \frac{\gamma(t+t^3)}{\sqrt{n}} \exp\left(-0.22(1-f\alpha^2)\right) \quad (4.6)$$

for $0 \leqq t \leqq \frac{(1-f\alpha^2)\sqrt{n}}{10\gamma}$.

By using (4.6) in Esséens inequality (Feller [4], p. 533)

$$|F(x)-G(x)| \leqq \frac{2}{\pi} \int_0^T \left|\frac{f(t)-g(t)}{t}\right| dt + \frac{24}{\pi T} \sup_x |G'(x)|. \quad (4.7)$$

We now easily arrive in the following theorem, which is the main result of this paper.

Theorem. *Let* $X_1, X_2, \ldots, X_N$ *be independent random variables and let* S_n *be the sum of* n *of them chosen at random. If* $EX_k=\mu_k$, $EX_k^2=\beta_k$, $E|X_k|^3=\gamma_k$, *where* $\sum_{k=1}^{N} \mu_k=0$, $\frac{1}{N}\sum_{k=1}^{N} \beta_k=1$, $\alpha^2=\frac{1}{N}\sum_{k=1}^{N} \mu_k^2$ *and* $\gamma=\max_{1\leqq k\leqq N} \gamma_k$, *then*

$$\left|P\left(\frac{S_n}{\sqrt{n(1-f\alpha^2)}} \leqq x\right)-\Phi(x)\right| \leqq \frac{60\gamma}{\sqrt{n}(1-f\alpha^2)^{\frac{3}{2}}}$$

where $\Phi(x)=(2\pi)^{-\frac{1}{2}} \int_{-\infty}^{x} \exp(-y^{\frac{2}{2}})\, dy$ is the normalized Gaussian distribution function and $f=\frac{n}{N}$.

Remark 1. The constant 60 is by no means the smallest one which satisfies this inequality. The value is a consequent of the way these calculations have been made and especially of the number 10 chosen in (3.1).

Remark 2. If $n=N$, S_n is the sum of N independent random variables. If in this case the variables have the same distribution, the remainder term in Theorem agrees with the one obtained by Esséen [3] (cf. Feller [4], p. 542).

Remark 3. If the set of variables $(X_1, X_2, \ldots, X_N)$ does not satisfy the normalizing conditions (1.1) and (1.2), we can easily obtain a new set $(X'_1, X'_2, \ldots, X'_N)$ which does satisfy (1.1) and (1.2), by a linear transformation. Application of the result of the theorem to this new set of variables gives, in terms of the original variables

$$\left|P\left(\frac{S_n - n\mu}{\sqrt{n\left[\frac{1}{N}\sum_{k=1}^{N}\sigma_k^2 + \frac{1-f}{N}\sum_{k=1}^{N}(\mu_k-\mu)^2\right]}} \leqq x\right) - \Phi(x)\right|$$

$$\leqq \frac{60}{\sqrt{n}} \frac{\max_{1\leqq k\leqq N} E|X_k-\mu|^3}{\left[\frac{1}{N}\sum_k \sigma_k^2 + \frac{1-f}{N}\sum_k(\mu_k-\mu)^2\right]^{\frac{3}{2}}}$$

where $\mu_k = EX_k$, $\mu = \frac{1}{N}\sum_{k=1}^{N}\mu_k$ and $\sigma_k^2 = \operatorname{Var} X_k$.

5. We will now indicate how this theoretical result may be used when estimating the mean μ of a finite universe by two-stage sampling.

Let the universe consist of N primary units $P_1, P_2, \ldots, P_N$, each of which consists of $M_1, M_2, \ldots, M_N$ secondary units respectively. Every secondary unit in P_k is characterized by a real number a_{kj}, $1\leqq j\leqq M_k$, $1\leqq k\leqq N$. The object of the statistical experiment is to estimate the mean

$$\mu = \frac{1}{M}\sum_{k=1}^{N}\sum_{j=1}^{M_k} a_{kj}, \quad \text{where} \quad M = \sum_{k=1}^{N} M_k$$

is the total number of secondary units in the universe.

Let now Z_k be the sum of the numbers a_{kj} obtained by a random selection of n_k secondary units out of P_k, $1\leqq k\leqq N$. If $m_k = \frac{1}{M_k}\sum_{j=1}^{M_k} a_{kj}$ is the mean and $s_k^2 = \frac{1}{M_k-1}\sum_{j=1}^{M_k}(a_{kj}-m_k)^2$ is the variance of P_k then $EZ_k = n_k m_k$ and $\operatorname{Var} Z_k = \frac{n_k(M_k-n_k)}{M_k} s_k^2$.

We now put $X_k = \frac{N}{M}\frac{M_k}{n_k} Z_k$ and apply the Theorem to the set $(X_1, X_2, \ldots, X_N)$. We have $\mu_k = EX_k = \frac{N}{M} M_k m_k$ and $\sigma_k^2 = \operatorname{Var} X_k = \frac{N^2}{M^2}\frac{M_k(M_k-n_k)}{n_k} s_k^2$. If S_n is the sum of n X_k's chosen at random, then $ES_n = \frac{n}{N}\sum EX_k = n\mu$, so $\mu^* = \frac{S_n}{n}$ is an unbiased estimate of μ, which, according to our theorem, is approximatively Gaussian with mean μ and variance

$$\frac{1}{nN}\left[\sum_{k=1}^{N}\frac{N^2}{M^2}\frac{M_k-n_k}{n_k} s_k^2 + (1-f)\sum_{k=1}^{N}\left(\frac{NM_k m_k}{M}-\mu\right)^2\right].$$

I am indebted to B. Rosén for correcting an error in a previous version of this paper.

References

1. Bikelis, A.: On the estimation of the remainder term in the central limit theorem for samples from finite populations (In Russian). Studia Sci. Math. Hungar. **4**, 345–354 (1969).
2. Erdös, P. and Rényi, A.: On the central limit theorem for samples from a finite population. Publ. Math. Inst. Hungar. Acad. Sci., Ser. A, **4**, 49–61 (1959).
3. Essén, C.-G.: Fourier analysis of distribution functions. Acta Math. **77**, 1–125 (1944).
4. Feller, W.: An Introduction to Probability Theory and Its Applications, Vol. II, Second Ed. New York: Wiley 1970.

Bengt von Bahr
Institute of Mathematical Statistics
and Actuarial Mathematics
University of Stockholm
Box 6701
S-113 85 Stockholm, Sweden

(Received November 13, 1971)